The Misspecification of the Covariance Structures in Multilevel Models for Single-Case Data: A Monte Carlo Simulation Study

Mariola Moeyaert and Maaike Ugille

University of Leuven, Belgium

John M. Ferron

University of South Florida, Tampa

S. Natasha Beretvas

University of Texas, Austin

Wim Van den Noortgate

Katholieke Universiteit Leuven, Belgium

Author Note

Mariola Moeyaert, Faculty of Psychology and Educational Sciences, Katholieke Universiteit Leuven, Belgium; Maaike Ugille, Faculty of Psychology and Educational Sciences, Katholieke Universiteit Leuven, Belgium; John M. Ferron, Department of Educational Measurement and Research, University of South Florida. S. Natasha Beretvas, Department of Educational Psychology, University of Texas; Wim Van den Noortgate, Faculty of Psychology and Educational Sciences, ITEC-iMinds Kortrijk, Katholieke Universiteit Leuven, Belgium.

This research is funded by the Research Foundations - Flanders, Grant number 12H1315N. The opinions expressed are those of the authors and do not represent views of the Institute or the Research Foundation - Flanders.

For the simulations we used the infrastructure of the Flemish Supercomputer Center, financed by the Department of Economy, Science and innovation – Flemish Government and the Hercules Foundation.

Correspondence concerning this article can be addressed to Mariola Moeyaert, Faculty of Psychology and Educational Sciences, Katholieke Universiteit Leuven, Andreas Vesaliusstraat 2 - Box 3762, B-3000 Leuven, Belgium. Phone +32 16 326091 or +32 16 326201. Fax +32 16 326200.
E-mail Mariola.Moeyaert@ppw.kuleuven.be

Abstract

The impact of misspecifying covariance matrices at the second and the third levels of the three-level model on inferences regarding the overall treatment effect and (co)variances estimates is evaluated by means of a simulation study and an empirical illustration is given. The results indicate that ignoring an existing covariance has no effect on the treatment effect estimate. In addition, the between-case variance estimates are unbiased and well estimated either when covariance is modeled or ignored. If the research interest lies in the between-study variance estimate, including at least 30 studies is warranted to get unbiased and precise estimates. Modeling covariance does not result in less biased and more precise between-study variance estimates as the between-study covariance estimate is extremely biased. As a consequence, single-case researchers can use either the model ignoring or modeling covariance when the research interest lies in the overall treatment effect estimate and/or the between-case variance. In addition, when the research interest lies on the between-case covariance, the model including covariance results in unbiased between-case variance estimates. The three-level model appears to be less appropriate to estimate the between-study variance if less than 30 studies are included.

Keywords: Multilevel modeling; multiple-baseline across cases designs; covariance misspecification; Monte Carlo simulation study

The Consequences of Misspecifying Covariance Structures in Multilevel Models for Single-Case Data: A Monte Carlo Simulation Study

Single-case experimental designs make important contributions to the field of educational research (National Research council, 2002; Odom, Brantlinger, Gersten, Horner, Thompson, & Harris, 2005). For instance, this kind of design can be applied to evaluate specific interventions to reduce challenging behavior in persons with intellectual disabilities or to search for strategies for persons with learning disabilities. Although single-case designs (SCDs) are increasingly popular (Kazdin, 2011), the quantitative analysis of study results obtained with this kind of design is still developing (Kratochwill et al., 2010). The results of a SCD study investigating the effect of an intervention are especially informative for the specific case under investigation, but it is hard to generalize conclusions to other cases. To investigate generalizability of the SCD results across cases, one can collect information for several cases, as is done in the multiple-baseline design (MBD) across cases. In this type of design, an AB phase design is implemented simultaneously to different cases, while the start point of the treatment is staggered (as in Figure 1) across the cases (Ferron & Scott 2005; Onghena, 2005; Onghena & Edgington, 2005).

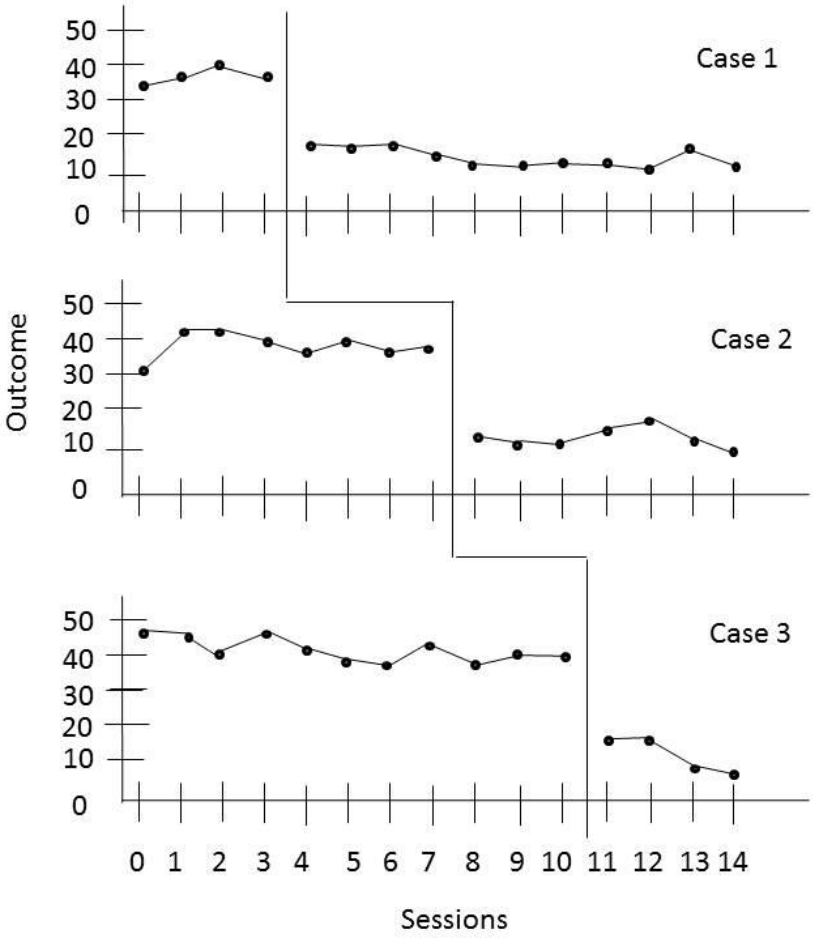

Figure 1. Graphical display multiple-baseline design across cases using hypothetical data. The start of the treatment is after session 3, session 7, and session 10 for Case 1, Case 2, and Case 3 respectively.

The MBD is growing in popularity because external events, which are random unexpected events influencing the outcome scores, can be disentangled from treatment effects. These external events might affect the outcome scores of several cases at the same time, while treatment effects are expected to occur immediately after the treatment starting point which is case-specific (Barlow, Nock & Hersen, 2009; Kinugasa, Cerin, & Hooper, 2004; Koehler & Levin, 2000).

To combine multiple cases' data, multilevel models can be used. Multilevel models are extensions of linear models and make it possible to synthesize treatment effects across cases and studies. When combining SCD data from several MBD studies, a three-level hierarchical structure can be modeled: measurement occasions (i.e., first level units) are nested within cases

(i.e., second level units), which in turn are nested within studies (i.e., third level units). For example, consider K studies (k = 0, 1,…, K), with J_k cases in study k (j = 0, 1,…J_k), and I_{jk} measurements for case j from study k (i = 0, 1,…I_{jk}). At level one, the continuous response variable can be modeled, for instance, using the following regression equation:

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk}D_{ijk} + e_{ijk} \text{ with } e_{ijk} \sim N(0, \sigma_e^2) \quad (1)$$

and the errors, the e_{ijk}'s, are assumed to be independently, identically, and normally distributed. The score on the continuous dependent variable on measurement occasion i for case j from study k (Y_{ijk}) depends on a binary coded treatment indicator (D_{ijk}), indicating whether measurement occasion i from case j within study k belongs to the baseline phase (D_{ijk} = 0) or the treatment phase (D_{ijk} = 1).

Equation 1, regressing Y_{ijk} on D_{ijk} contains two coefficients: β_{0jk} is the intercept and indicates the expected baseline level, and β_{1jk} refers to the treatment effect (i.e., the difference between the estimated outcome score under the treatment phase and the estimated outcome score under the baseline phase). SCD researchers are mainly interested in β_{1jk} because this coefficient provides information about the change associated with the introduction of the treatment.

At the second level, the variation across cases can be modeled as follows:

$$\begin{cases} \beta_{0jk} = \theta_{00k} + u_{0jk} \\ \beta_{1jk} = \theta_{10k} + u_{1jk} \end{cases} \text{ with } \begin{bmatrix} u_{0jk} \\ u_{1jk} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u_0}^2 & \sigma_{u_0u_1} \\ \sigma_{u_1u_0} & \sigma_{u_1}^2 \end{bmatrix} \right) \quad (2)$$

These equations indicate that the β coefficients from Equation 1 randomly vary across cases, around study-specific means, the θ coefficients. The coefficients along the diagonal of the covariance matrix, $\sigma_{u_0}^2$, $\sigma_{u_1}^2$, indicate the between-case variance in the intercept, and the treatment effect, respectively. The off-diagonal coefficient represents the covariance. $\sigma_{u_0u_1}$ indicates the covariance between the intercept and the treatment effect. It is for instance

reasonable to expect that participants with a higher baseline level will benefit less from the treatment and vice versa.

At the third level, potential variability in the study-specific regression coefficients from the second level equations, the θ coefficients, is modeled. In the fullest model, the θ coefficients each equal an overall estimate across studies, indicated by the γ coefficients, and a random deviation from this average:

$$\begin{cases}\theta_{00k} = \gamma_{000} + v_{00k} \\ \theta_{10k} = \gamma_{100} + v_{10k}\end{cases} \text{ with } \begin{bmatrix}v_{00k} \\ v_{10k}\end{bmatrix} \sim N\left(\begin{bmatrix}0 \\ 0\end{bmatrix}, \begin{bmatrix}\sigma^2_{v_0} & \sigma_{v_0 v_1} \\ \sigma_{v_1 v_0} & \sigma^2_{v_1}\end{bmatrix}\right) \quad (3)$$

Multilevel modeling entails the advantage that an overall treatment effect can be estimated, as well as variation between studies and cases in the treatment effect, or study- and case-specific treatment effects. Another major advantage of this multilevel approach is its flexibility. For instance, the model can be extended by including (additional) predictors at each level (e.g., time predictor at the first level, gender as case-specific predictor and average age as a study-specific predictor). Moreover, a specific structure for the variances and covariances at either level can be specified.

Previous research indicates that multilevel modeling works appropriately to combine unstandardized (Owens & Ferron, 2012; Moeyaert, ugille, Ferron, Beretvas, & Van den Noortgate, 2013a) and standardized (Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2013b) SCD data across cases and studies. Estimation of the three-level model for SCD data was investigated, by evaluating the estimates of the overall treatment effect, γ_{100}, and of the between-case and between-study variance of the overall treatment effect. However, in previous studies, the between-case residuals were each assumed to be independently, identically, and normally distributed with mean zero and homogeneous variance, and thus a diagonal covariance structure was assumed at level-2. This might be an over-simplification of the between-case covariance structure. A non-zero covariance between residuals at level 2 seems reasonable, for instance, when due to a ceiling effect the treatment effect is expected to be smaller for cases

with an already high baseline level. In addition, these simulation studies made the same assumptions about the between-study residuals but again an unstructured covariance matrix, which allows the level-3 residuals to covary, may be more reasonable than a diagonal covariance structure. To date, no research has focused on the consequences of ignoring truly non-zero covariances in the context of multilevel modeling of SCD data. In most multilevel modeling software, the default option is to estimate an unconstrained covariance matrix for the random effects. However, given that there are two coefficients included in the level-1 regression equation, a total of 7 random effects covariance parameters have to be estimated, which complicates estimation especially in scenarios with small sample sizes and possible covariance values that are close to zero. Therefore there is a need to investigate if estimation of the multilevel model is robust to covariance matrix misspecification. If estimation is reasonably robust, then modeling of a simplified covariance matrix can be recommended for future studies using the three-level model in the context of SCD data.

While previous research in the context of multilevel modeling of SCD data mainly focused on the fixed effect estimates, there is some methodological research that has focused on specification of the residuals' covariance matrix in contexts other than SCDs. For example, Singer and Willett (2003) argue that ignoring a covariance in a multilevel model in general may bias the estimation of the standard errors of the overall regression coefficients. This will in turn lead to distorted Type I error rates when testing the statistical significance regression coefficients and will affect estimation of the confidence intervals for the effects of interest. Kwok, West and Green (2007) investigated by means of a simulation study the misspecification of the within-case covariance structure in multiwave longitudinal multilevel models and found that the misspecification has a substantial impact on the variance estimates. Work by Berkhof and Kampen (2004) examine the effect of omitting a random coefficient in the multilevel models in general on the estimated variance components and the estimated variance of the treatment effects. They found that the consequences depend on the between-unit variance

proportions. Another study, by Van den Noortgate and Onghena (2005), investigated the effects of ignoring a level from a four level model (in the area of school effectiveness research) on the parameter estimates and standard errors. They found that the variance estimate of the ignored level is divided between the other levels and estimates of the standard errors of the fixed effects and the random components may change.

Specifically for SCD data, the effects of level-1 residuals' covariance misspecification have been studied before for a two-level model (Ferron, Bell, Hess, Rendina-Gobioff, 2009). In SCDs, it is reasonable that an external variable that influences an observation at a certain moment, also affects succeeding observations. This means that errors from succeeding occasions tend to be more alike than errors of occasions further in time (Kromrey & Foster-Johnson, 1996). Ferron et al. (2009) found that not modeling autocorrelation in a two-level analysis of SCD data results in too small coverage proportions of the 95% confidence intervals and positively biased variance estimates. This same pattern of results also apply when level-1 residuals' autocorrelation is not modeled for the three-level model (Petit-Bois, Baek, & Ferron, 2012). Level-2 and level-3 covariance misspecification issues in the SCD three-level modeling framework have not yet been investigated. The main focus of this paper is to examine the consequences of level-2 and level-3 covariance matrix misspecification which should provide a more complete understanding of misspecification issues in contexts of three-level modeling of SCD data.

Simulation Study

We conducted two simulation studies to evaluate estimation of the three-level model when freely estimating covariances between pairs of residuals at levels two and three of the three-level model. It might be possible to mathematically derive large-sample approximations of the estimated standard errors of the treatment effects. However, in the context of multilevel modeling of SCD data, researchers deal with very small sample sizes which violate asymptotic assumptions upon which the algebraic derivations would be based. Thus, we exclusively rely

on simulation studies to empirically examine estimation of model parameters and standard errors under the realistic sample size values that are typically encountered in applied SCD research in educational and social sciences.

We simulated raw MBD across cases data using the three-level model, which is obtained by combining Equations 1 through 3:

$$Y_{ijk} = y_{000} + v_{00k} + u_{0jk} + (\gamma_{100} + v_{20k} + u_{2jk})D_{ijk} + e_{ijk} \quad (4)$$

To estimate the three-level model parameters, the restricted maximum likelihood procedure in SAS 9.3 PROC MIXED was used (Littell, Milliken, Stroup, & Wolfinger, 2006). The Kenward-Roger method was used to estimate the degrees of freedom because this method is accurate and appropriate for unbalanced designs and has the additional advantage that it corrects for small sample sizes in the variance estimation (Kenward & Roger, 1997).

The criteria used to evaluate the overall treatment effect estimate across cases and across studies using the three-level analysis included the (relative) bias, the mean squared error (*MSE),* the relative standard error bias, and the coverage proportion of the 95% confidence intervals of the overall treatment effect (i.e., $\hat{\gamma}_{100}$). In contrast to previous simulation studies, we also evaluate the performance of the three-level modeling technique to estimate the variance components estimates (i.e., the between-case and the between-study variance of the treatment effect) and whether ignoring the covariance in the dataset has effects on the between-case variance and the between-study variance of the treatment effect in terms of the accuracy of the estimate (i.e., relative bias) and the precision of the estimate (i.e., *MSE*).

The purpose of this study is to evaluate the performance of two multilevel models to estimate the overall treatment effect across cases and across studies and the between-case variance and the between-study variance of the treatment effect; one in which no covariance is estimated (i.e., Analysis Model 1) and one in which covariance is estimated (i.e., Analysis

Model 2). In order to evaluate these models in a systematic way, we conducted two simulation studies.

In Simulation Study 1, we generated data without covariance between the baseline level and the treatment effect at both level 2 and level 3 and estimated the overall treatment effect and the variance components using Analysis Model 1 and Analysis Model 2. In Simulation Study 2, we generated covariance between the baseline level and the treatment effect on both level 2 and level 3 (taking on small and large values and crossing these values) and used again Analysis Model 1 and Analysis Model 2 to estimate the model parameters of interest. We are especially interested whether Analysis Model 2 results in more accurate and precise variance components estimates and less biased estimates of the standard error of the overall treatment effect.

For the first Monte Carlo simulation study, we varied six design conditions, namely the treatment effect (i.e., γ_{100}), the number of units at the three levels (i.e., the number of measurements at the first level, *I*, the number of cases at the second level, *J*, and the number of studies at the third level, *K*), and the between-case, and between-study variability (i.e., $\sigma^2_{u_1}$ and $\sigma^2_{u_1}$ respectively). For Simulation Study 2, two additional design conditions were varied, namely the covariance between the baseline level and the treatment effect on both level 2 (i.e., $\sigma_{u_0u_1}$) and level 3 (i.e., $\sigma_{v_0v_1}$). The conditions were fully crossed and 1,000 datasets were generated for each condition.

In order to identify values for the design conditions that are authentic for SCD data encountered in the area of educational research, we re-analyzed published meta-analyses of SCD studies (Denis, Van den Noortgate, & Maes, 2011; Heyvaert, Maes, Van den Noortgate, Kuppens, & Onghena, 2011; Kokina & Kern, 2010; Shogren, Fagella-Luby, Bae, & Wehmeyer, 2004; Wang, Cui, & Parrila, 2011). Based on the re-analysis, a median overall treatment effect of 2 was observed. In addition, we were also interested whether a zero treatment effect can be

recovered using the three-level model. Therefore, γ_{100}, was given a value of 0 or 2. As a consequence, data were generated such that the treatment causes an increase in outcome scores (e.g., the score on a math test). The regression coefficient indicating the baseline level (i.e., γ_{000}) was not varied (and set to 0), because the focus of the current study is on the overall treatment effect estimate.

The number of units at the three levels was varied. The highest level represents the study level. Each of these studies includes a limited number of cases (i.e., which represents the second level units) and in turn, in each case a number of measurement occasions is clustered (i.e., which is the lowest level). The number of simulated studies was set to 10 or 30 (K = 10 or 30) based on a review of social science single-case meta-analysis by Farmer, Owens, Ferron, and Allsopp (2010). They showed that 60% of the meta-analysis included less than 30 studies. We choose to test whether the multilevel model already works appropriate including lower limits for the number of studies. The number of cases simulated per study took on a value of 3 or 7 (J = 3 or 7). These values are chosen based on the re-analysis of meta-analysis of single-cases (Denis et al., 2011; Heyvaert et al., 2011; Kokina & Kern, 2010; Shogren et al., 2004; Wang et al., 2011), on guidelines that MBDs have at least 3 baselines (Barlow, Nock & Hersen, 2009), on a review of MBD studies (Ferron, Farmer & Owens, 2010) illustrating that studies have between 3 to 10 cases, on a survey of Shadish and Sullivan (2011) stating that the number of cases per study fell between 1 and 13 with a median of 3, and a review of Farmer et al. (2010) demonstrating that 93% of studies included 7 or less than 7 cases. The generated number of measurement within a case varied and consisted of 20 or 40 measurement occasions (I = 20 or 40). Ferron et al. (2010) found a median of 24 measurements and Shadish and Sullivan (2011) found a similar value (i.e., 20) and documented that 90.6% of the cases had fewer than 49 data points. As we choose an MBD across cases' designs, we staggered the introduction of the intervention across cases within studies. The staggering is a function of the total number of measurements and cases (see Table 1) within an MBD study.

Table 1

Staggering of the Intervention's Start Point as a Function of the Number of Cases (*J*) and Measurement Occasions (*I*)

J	*j*	Start of treatment	
		I = 20	*I* = 40
3	1	7	11
	2	10	18
	3	12	24
7	1	7	11
	2	9	15
	3	9	15
	4	11	21
	5	13	27
	6	13	27
	7	15	31

Note. The data in the simulation study were generated using an MBD across cases design and therefore the introduction of the intervention is staggered across cases within an MBD study. *J* indicates the number of cases (*j* = 1 to *J*) and *I* indicates the number of measurement occasions in one case. The cells of the last two columns represent the time at which the intervention is started.

The within-case variance was generated with a variance of one and assumed to be homogeneous across phases. Level-2 and level-3 errors were generated from a normal distribution using the RANNOR random number generator in SAS. The between-case covariance matrix (i.e., Σ_u) was manipulated representing conditions with relatively small and relatively large amounts of between-case variance. In Simulation Study 1, the covariance was set to zero (at both the case and study level) and therefore in Simulation Study 1, Σ_u is a diagonal matric, $\Sigma_u = diag(\sigma^2_{u_0}, \sigma^2_{u_1})$. Our re-analyses of meta-analyses (Denis et al., 2011; Kokina & Kern, 2010; Shogren et al., 2004) indicated that the variances are twice or eight times larger than the within-case variance and therefore we chose values of 2 or 8: $\Sigma_u = diag(\sigma^2_{u_0}, \sigma^2_{u_1})$ = diag (2 , 2) or $\Sigma_u = diag(\sigma^2_{u_0}, \sigma^2_{u_1})$ = diag (8 , 8). Again based on re-analyses of meta-analyses (Alen et al., 2009; Denis et al., 2011; Kokina & Kern, 2010; Shogren et al., 2004; Wang et al., 2011), we have chosen the same sets of values for the two diagonal elements of the between-study variance. In Simulation Study 2, covariance between the intercept and the treatment effect were generated. Based on Alen et al. (2009); Denis et al. (2011); Kokina & Kern (2010); Shogren et al. (2004); Wang et al. (2011), negative covariances at level 2 and level

3 were generated, indicating that a rather large baseline level results in a rather small treatment effect estimate. In order to define the values for the covariance, we calculated the correlation between the baseline level and the treatment effect using the dataset obtained by the re-analyses of meta-analyses (Denis et al., 2011; Kokina & Kern, 2010; Shogren et al., 2004). We found that the correlation between the baseline level and the treatment effect on both level 2 and level 3 ranged from -0.30 to -0.70 and therefore we choose to include this lower and upper limit. Using the correlation and the standard deviation of the baseline level and treatment (i.e., root of the variance), we can calculate the covariance between the baseline level and the treatment effect. For instance, for the covariance at the third level, we can calculate the between-study covariance by transforming Equation 5:

$$r_{v_0 v_1} = \frac{\sigma_{v_0 v_1}}{\sigma_{v_0}\,\sigma_{v_0}} \tag{5}$$

For instance, if the correlation equals -.30 and the variances are set on 2 or 8 , then the covariance equals -0.60 or -2.40 respectively. On the other hand , if the correlation equals -.70 and the variances are set on 2 or 8 , then the covariance equals -1.40 or -5.60 respectively. As a matter of clarity, the conditions for Simulation Study 1 and Simulation Study 2 are presented in Table 2.

Table 2

Overview Design Conditions for Simulation Study 1 (i.e., no Covariance is Generated) and Simulation Study 2 (i.e., Covariance is Generated)

	Design Condition	Notation	Design Condition Values	
			Simulation Study 1	Simulation Study 2
1	Treatment effect	γ_{100}	γ_{100} = 0 or 2	γ_{100} = 0 or 2
2	Level-1 sample size (i.e., number of measurements)	I	I = 20 or 40	I = 20 or 40
3	Level-2 sample size (i.e., number of cases)	J	J = 4 or 7	J = 4 or 7
4	Level-3 sample size (i.e., number of studies)	K	K = 10 or 30	K = 10 or 30
5	Between-case variance - Intercept - Treatment effect	$\sigma^2_{u_0}$ $\sigma^2_{u_1}$	$\sigma^2_{u_0}/\sigma^2_{u_1}$ = 2 or 8	$\sigma^2_{u_0}/\sigma^2_{u_1}$ = 2 or 8
6	Between-study variance - Intercept - Treatment	$\sigma^2_{v_0}$ $\sigma^2_{v_1}$	$\sigma^2_{v_0}$ / $\sigma^2_{v_1}$ = 2 or 8	$\sigma^2_{v_0}$ / $\sigma^2_{v_1}$ = 2 or 8
7	Between-case covariance	$\sigma_{u_0u_1}$	/	If $\sigma^2_{u_0}$ / $\sigma^2_{u_1}$ = 2 then $\sigma_{u_0u_1}$ = -0.60 or -1.40 If $\sigma^2_{u_0}$ / $\sigma^2_{u_1}$ = 8 then $\sigma_{u_0u_1}$ = -2.40 or -5.60
8	Between-study covariance	$\sigma_{v_0v_1}$	/	If $\sigma^2_{v_0}$ / $\sigma^2_{v_1}$ = 2 then $\sigma_{v_0v_1}$ = -0.60 or -1.40 If $\sigma^2_{v_0}$ / $\sigma^2_{v_1}$ = 8 then $\sigma_{v_0v_1}$ = -2.40 or -5.60

Note. The values for the covariance at the second level ($\sigma_{u_0u_1}$) and third level ($\sigma_{v_0v_1}$) can take on two values representing a small correlation (r = -.30) and a large correlation (r = -.70) between baseline level and treatment level. The conditions were fully crossed and 1,000 datasets were generated for each condition.

As a consequence, we examined a total of $2^6 = 64$ conditions in Simulation Study 1 and $2^8 = 256$ conditions in Simulation Study 2. For each condition we simulated 1,000 datasets resulting in 64,000 and 256,000 datasets to analyze for Simulation Study 1 and Simulation Study 2 respectively. In Simulation Study 1, no covariance was generated in contrast to Simulation Study 2. In both simulation studies, the generated datasets were analyzed twice (i.e., Analysis Model 1 vs Analysis Model 2) using a three-level multilevel model with restricted maximum likelihood estimation via the MIXED procedure in SAS. In Analysis Model 1, the covariances at level 2 and level 3 between the baseline level and the treatment effect were constrained to zero, whereas in Analysis Model 2, the covariance components on both levels were freely estimated.

We expect no differences between Model 1 and Model 2 in Simulation Study 1 (in which no covariance is generated). In Simulation Study 2, we expect biased standard error estimates of the treatment effect and less accurate and precise variance components when Analysis Model 1 is used (i.e., covariance is generated but ignored in the analysis).

Results

First the results of Simulation Study 1 are presented in which no covariance between baseline level and treatment effect was generated followed by a second simulation study in which covariance was simulated. We used two-way ANOVAs (PROC GLM in SAS) as preliminary analyses to explore whether the bias, the *MSE*, the relative standard error bias and the coverage proportion of the 95% confidence interval of the overall treatment effect estimate depend on the analysis model (i.e. Analysis Model 1 vs Analysis Model 2) or certain design conditions. As indicted earlier (and displayed in Table 2), 6 design conditions (i.e., γ_{200}, *K*, *I*, *J*, $\sigma^2_{u_2}$, $\sigma^2_{v_2}$) are evaluated in Simulation Study 1, whereas 8 design conditions are explored in Simulation Study 2 (i.e., γ_{200}, *K*, *I*, *J*, $\sigma^2_{u_2}$, $\sigma^2_{v_2}$, $\sigma_{u_0u_1}$, $\sigma_{v_0v_1}$). We did not only look at statistically significant main effects and interaction effects ($p < .001$), but we also calculated the eta squares as effect sizes to evaluate whether the estimated main effects and/or interaction effects have a

rather small ($\leq$ .02), medium (.03-.25) or large ($\geq$.26) effect (Cohen, 1988). In addition, we are interested in how accurate (evaluated by the relative bias) and precise (evaluated by the *MSE*) the variance components are estimated using the two different analysis models. The relative bias of the overall treatment effect estimate cannot be calculated when the population value is set to 0 and therefore we evaluate the absolute bias for the fixed effect estimates (i.e., difference between the estimated value and the true population effect). For the (co)variance estimate, it is possible to calculate the relative bias (i.e., absolute bias divided by the population value) as the population values are nonzero. Whenever it is possible it is preferred to evaluate the relative bias rather than the absolute bias as this takes into account the magnitude of the population effect. For the relative bias estimates and the relative standard error bias estimates, we considered .05 and .10 respectively as substantial (Hoogland & Boomsma, 1998).

Simulation Study 1

Bias and *MSE* of the overall treatment effect estimate. We first look at the absolute bias, which is generally defined as the mean difference between the estimated values and the population value. From the ANOVA, we conclude that the analysis model has no statistically significant and large effect on the bias [$F(1, 128032) = 0.13$, $p = .720$, $\hat{\eta}^2 < .0001$] and the *MSE* [$F(1, 128032) = 0.690$, $p = .41$, $\hat{\eta}^2 < .0001$] of the overall treatment effect estimate. As indicated in Table 3, only the amount of between-case variance has a statistically significant, but small effect on the relative bias [$F(1, 128032) = 13.01$, $p < .001$, $\hat{\eta}^2 = .0001$). The absolute bias ranges from 5.660×10^{-4} to 0.0736 and the largest bias is observed in the conditions characterized by a small amount of studies ($K = 10$) and a large amount of between-study variance (i.e., $\sigma^2_{u_1} = 8$) as illustrated in Figure 2.

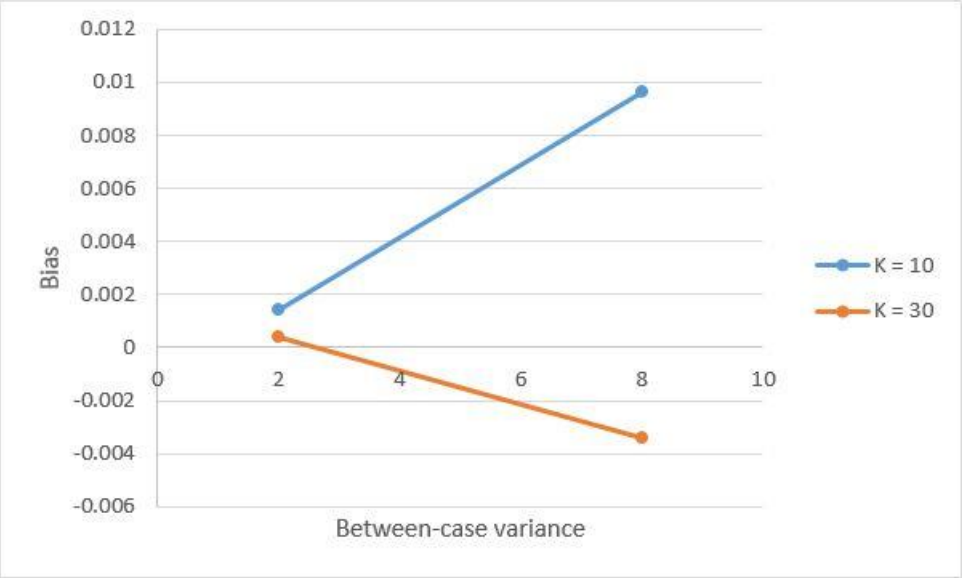

Figure 2. Graphical display of the influence of the between-case variance and the numher of studies on the bias of the overall treatment effect estimates in simulation study 1 (i.e., no covariance is generated*). K* indicates the number of studies. The graphical display is for a subset for conditions as similar patterns where found for the other combination of conditions. Analysis Model 1 (i.e., no covariance is estimated) is used, the overall baseline level is set to zero, the number of measurements and cases is set to 20 and 3 respectively, and the between-study variance equals 2.

In addition, we looked at the *MSE*, an important criterion indicating how precise the treatment effect is estimated. A small bias in treatment effect estimate does not necessarily imply a small *MSE*. It can be the case that the individual estimates vary a lot around the mean estimate. An interesting finding is that the number of cases [$F(1, 128032) = 272.15$, $p < .001$, $\hat{\eta}^2 = .0018$], the number of studies [$F(1, 128032) = 9987.14$, $p < .001$, $\hat{\eta}^2 = .0660$] and the amount of variance at level 2 [$F(1, 128032) = 541.48$, $p < .0001$, $\hat{\eta}^2 = .0036$] and level 3 [$F(1, 128032) = 9687.85$, $p < .0001$, $\hat{\eta}^2 = 0.0641$] of the multilevel model have a statistically significant effect on the *MSE* of the estimated treatment effect. However, only the between-study variance and the number of studies appear to have a medium effect as indicated by the $\hat{\eta}^2$ (Cohen, 1988). All the other design conditions and interactions have a small effect as indicated in Table 3. From Figure 3 we can deduce that the larger the number of level-3 units and the smaller the between-study variance, the smaller the *MSE*. The *MSE* ranges from 0.073 to 1.120

and the largest *MSE* is identified when 10 studies, 3 cases, 40 measurement occasions, a large amount of between-case variance (i.e., $\sigma^2_{u_1}$ = 8) and a large amount of between-study variance (i.e., $\sigma^2_{u_1}$ = 8) in combination with a zero treatment effect is included, independent of the analysis model.

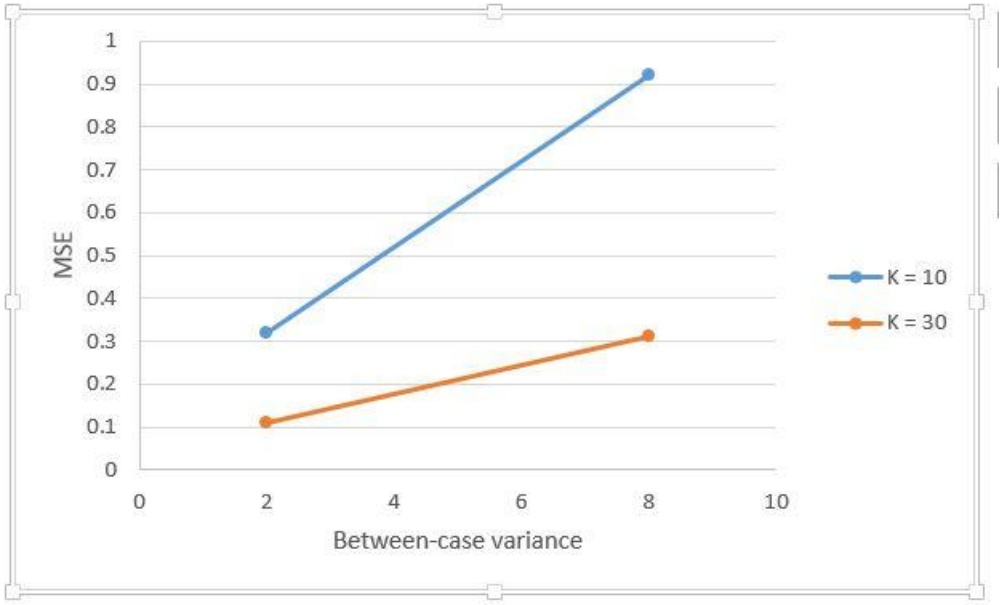

Figure 3. Graphical display of the influence of the between-case variance and the numher of studies on the mean squared error of the overall treatment effect estimates in Simulation Study 1 (i.e., no covariance is generated). *K* indicates the number of studies. The graphical display is for a subset for conditions as similar patterns where found for the other combination of conditions. Analysis Model 1 (i.e., no covariance is estimated) is used, the overall baseline level is set to zero, the number of measurements and cases is set to 20 and 3 respectively, and the between- study variance equals 2.

Table 3

Results Simulation Study 1 (i.e., no Covariance is Generated) using Analysis Model 1 (no Covariance is Estimated). Evaluation of the Main Effects and Interaction Effects of Simulation Conditions on the Relative Bias, Mean Squared Error, Standard Error and Coverage Proportion of the 95% Confidence Interval of the Treatment Effect

		Relative Bias			Mean Squared Error			Relative Standard Error Bias			Coverage Proportion 95% Confidence Interval		
Independent Variable	*df*	*F*	*p*	$\hat{\eta}^2$	*F*	*p*	$\hat{\eta}^2$	*F*	*p*	$\hat{\eta}^2$	*F*	*p*	$\hat{\eta}^2$
θ_{100}	1	0.02	.889	<.0001	1.22	.270	<.0001	4.21	.043	.0271	2.86	0.094	.0210
I	1	2.31	.128	<.0001	0.25	.616	<.0001	0.01	.928	.0001	0.40	0.527	.0030
J	1	0.27	.604	<.0001	272.16*	<.0001	.0018	1.50	.223	.0097	0.77	0.383	.0056
K	1	2.79	.095	<.0001	9987.46*	<.0001	.0660	11.90*	.001	.0765	8.53	0.004	.0626
$\sigma^2_{u_1}$	1	13.01*	<.0001	.0001	541.50*	<.0001	.0036	0.77	.381	.0050	0.07	0.795	.0005
$\sigma^2_{v_2}$	1	0.38	.539	<.0001	9688.15*	<.0001	.0641	1.86	.175	.0120	0.79	0.375	.0058
θ_{100}*I	2	0.13	.879	<.0001	0.69	.502	<.0001	0.01	.904	.0001	0.01	0.916	.0001
θ_{100}*J	1	0.35	.556	<.0001	2.05	.152	<.0001	3.72	.057	.0239	1.62	0.206	.0119
θ_{100}*K	1	6.10	.014	<.0001	3.46	.063	<.0001	5.81	.018	.0374	3.34	0.070	.0245
θ_{100}*$\sigma^2_{u_1}$	1	0.12	.725	<.0001	1.4	.237	<.0001	3.85	.053	.0247	0.03	0.874	.0002
θ_{100}*$\sigma^2_{v_1}$	1	2.05	.152	<.0001	0.73	.393	<.0001	1.11	.295	.0071	1.76	0.188	.0129
*I*J*	1	0.00	.968	<.0001	1.62	.203	<.0001	1.14	.287	.0074	0.69	0.408	.0051
*I*K*	1	4.07	.044	<.0001	2.09	.148	<.0001	0.02	.875	.0002	1.49	0.226	.0109
I$\sigma^2_{u_1}$	1	1.09	.296	<.0001	0.01	.933	<.0001	0.53	.468	.0034	0.80	0.374	.0059
I$\sigma^2_{v_1}$	1	0.38	.539	<.0001	0.04	.840	<.0001	1.44	.233	.0093	0.73	0.394	.0054
*J*K*	1	1.37	.241	<.0001	1.04	.309	.0005	3.81	.054	.0245	0.19	0.666	.0014
J$\sigma^2_{u_1}$	1	0.16	.690	<.0001	72.68*	<.0001	.0006	0.15	.695	.0010	0.22	0.641	.0016
J$\sigma^2_{v_1}$	1	1.78	.182	<.0001	94.18*	<.0001	<.0001	3.72	.057	.0239	1.08	0.302	.0079
K$\sigma^2_{u_1}$	1	2.36	.124	<.0001	2.63	.105	.0158	1.96	.165	.0126	3.54	0.063	.0260
K$\sigma^2_{v_1}$	1	3.81	.051	<.0001	2390.06*	<.0001	.0011	1.98	.162	.0127	1.01	0.318	.0074
$\sigma^2_{u_1}$*$\sigma^2_{v_1}$	1	0.00	.990	<.0001	161.32*	<.0001	<.0001	0.22	.644	.0014	1.45	0.231	.0107

Note. *$p < .001$. $\hat{\eta}^2$ reflects the magnitude of the main effects and/or interaction effects on the relative bias, mean squared error, relative standard error bias and coverage proportion of the 95% confidence interval. $\hat{\eta}^2$ can take on values between 0 and 1 (small: $\hat{\eta}^2 \leq .02$, medium: $\hat{\eta}^2$ = .03-.25, and large: $\hat{\eta}^2 \geq .26$, Cohen, 1988). θ_{100} = immediate treatment effect, *I*, *J*, and *K* = number of measurements, cases, and studies respectively. Interactions between two independent variables are indicated with a * in between the two independent variables (e.g., θ_{100}*I indicates an interaction between the immediate treatment effect and the number of measurements).

Standard error and coverage proportion of the 95% confidence interval of the overall treatment effect estimate. The standard errors of the treatment effect estimates are used to construct confidence intervals around the estimated treatment effect, $\hat{\gamma}_{100}$. The standard deviations of the effect estimates in a given condition can be used as an empirical approximation of the true standard error and therefore as a criterion to evaluate the standard error estimates. We looked at the relative standard error bias, which is the difference between the mean standard error estimate and the standard deviation of the estimate of the effect divided by the standard deviation of the estimate of $\hat{\gamma}_{100}$ (Hoogland & Boomsma, 1998). The estimated standard error is independent of the analysis model [$F(1, 44) = 0.81$, $p = .371$, $\hat{\eta}^2 = .0052$]. We found that the number of studies [$F(1, 44) = 11.9$, $p = < .001$, $\hat{\eta}^2 = .0765$] is the only design condition having a statistically significant medium influence on the relative standard error bias. Table 3 gives a complete overview of the effect of the other design conditions and interactions on the relative standard error bias. The values for the relative standard error biases are negative, which means that the standard error estimates are slightly smaller than expected. However, none of the conditions reports a relative standard error bias larger than 10% which is widely accepted as cut off score (Hoogland & Boomsma, 1998). From this, we conclude that the standard error of overall treatment effect is well estimated. In addition, the coverage proportion of the 95% confidence interval was calculated. The estimated coverage proportion ranges from .94 to .96 with an average across all condition of .95 which is what we expected. The coverage proportion of the 95% confidence intervals appears to be independent of the analysis model [$F(1, 104) = 0.37$, $p = .544$, $\hat{\eta}^2 = .0027$] and other design conditions (indicated by the rather low value for the $\hat{\eta}^2$ in Table 3).

Bias and *MSE* of the between-study variance of the treatment effect estimate. In addition to the treatment effect estimate, we were interested in how accurate (i.e., relative bias) and precise (i.e., *MSE*) the between-study variance of the treatment effect is estimated. The analysis model (Analysis Model 1 vs Analysis Model 2) has no statistically significant

large effect on the relative bias [$F(1, 128004) = 1.67$, $p = .196$, $\hat{\eta}^2 < .0001$] and the *MSE* [$F(1, 128004) = 0.75$, $p = .386$, $\hat{\eta}^2 < .0001$] of the estimated between-study variance. The between-study variance [$F(1, 128004) = 39.83$, $p < .001$, $\hat{\eta}^2 = .0003$] and the between-case variance [$F(1, 128004) = 26.41$, $p < .001$, $\hat{\eta}^2 = .0002$] have a statistically significant but small effect on the relative bias. No other design conditions or interactions appear to have a statistically significant or large effect on the relative bias of the between-study variance estimate as indicated in Table 4. The relative bias exceeds the cut off criterion of 5% set by Boomsma and Hoogland (1998) in almost all conditions except when 30 studies are included and the between-case and between-study variance have equal values (i.e., $\sigma^2_{u_2} = \sigma^2_{u_2} = 2$ or $\sigma^2_{u_2} = \sigma^2_{u_2} = 8$). The relative bias ranges from 0.761% to 303.927%. The problematic large values are obtained in conditions characterized by an unequal amount of between-study and between-case variance as illustrated in Figure 4.

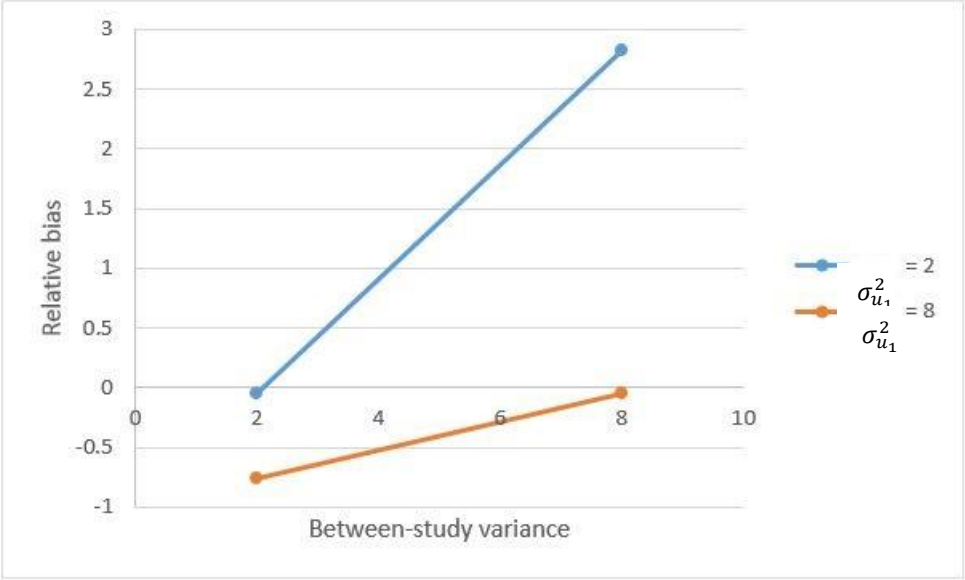

Figure 4. Graphical display of the influence of the between-case variance and the between-study variance on the relative bias of the between-study variance estimate in Simulation Study 1 (i.e., no covariance is generated). $\sigma^2_{u_1}$ indicates the between-case variance. The graphical display is for a subset for conditions as similar patterns where found for the other combination of conditions. Analysis Model 1 (i.e., no covariance is estimated) is used, the overall baseline

level is set to zero, the number of measurements, cases and studies is set to 20, 3, and 10 respectively.

Table 4 also displays the effects of the design conditions and interactions on the *MSE* of the between-study variance. From this table it is clear that no statistically significant large effects are identified. The values for the *MSE* range from 0.166 to 41.613. The largest *MSE* values correspond to the condition in which the largest relative bias was detected (i.e., unequal amount of between-case and between-study variance).

Table 4

Results Simulation Study 1 (i.e., no Covariance is Generated) using Analysis Model 1 (no Covariance is Estimated). Evaluation of the Main Effects and Interaction Effects of Simulation Conditions on the Relative Bias and Mean Squared Error of the Between-Study Variance and Between-Case Variance Estimate

		Between-Study Variance Estimate						Between-Case Variance Estimate					
		Relative Bias			Mean Squared Error			Relative Bias			Mean Squared Error		
Independent Variable	*df*	*F*	*p*	$\hat{\eta}^2$	*F*	*p*	$\hat{\eta}^2$	*F*	*p*	$\hat{\eta}^2$	*F*	*p*	$\hat{\eta}^2$
θ_{100}	1	0.48	.487	< .0001	0.02	.883	< .0001	1.89	.1690	< .0001	1.08	.2994	< .0001
I	1	1.10	.294	< .0001	0.91	.340	< .0001	2.25	.1336	< .0001	1.15	.2832	< .0001
J	1	1.27	.260	< .0001	0.97	.325	< .0001	0.32	.5688	< .0001	0.27	.6040	< .0001
K	1	3.31	.069	< .0001	3.84	.050	< .0001	2.42	.1201	< .0001	1.57	.2104	< .0001
$\sigma^2_{u_1}$	1	26.41*	<.0001	.0002	0.13	.721	< .0001	0.26	.6103	< .0001	2.22	.1360	< .0001
$\sigma^2_{v_2}$	1	39.83*	<.0001	.0003	3.84	.050	< .0001	2.54	.1107	< .0001	1.57	.2105	< .0001
θ_{100}*I	2	2.12	.146	< .0001	1.33	.248	< .0001	3.88	.0488	< .0001	3.16	.0757	< .0001
θ_{100}*J	1	0.52	.471	< .0001	0.61	.434	< .0001	0.01	.9409	< .0001	1.49	.2219	< .0001
θ_{100}*K	1	0.44	.509	< .0001	0.02	.885	< .0001	0.72	.3963	< .0001	0.23	.6330	< .0001
θ_{100}*$\sigma^2_{u_1}$	1	0.84	.360	< .0001	1.72	.190	< .0001	0.04	.8390	< .0001	0.51	.4738	< .0001
θ_{100}*$\sigma^2_{v_1}$	1	0.46	.498	< .0001	0.02	.885	< .0001	0.71	.4002	< .0001	0.23	.6327	< .0001
*I*J*	1	0.14	.709	< .0001	0.00	.980	< .0001	0.00	.9558	< .0001	1.58	.2087	< .0001
*I*K*	1	1.06	.303	< .0001	0.90	.342	< .0001	0.95	.3289	< .0001	0.26	.6080	< .0001
I$\sigma^2_{u_1}$	1	0.35	.556	< .0001	0.26	.613	< .0001	0.01	.9436	< .0001	0.56	.4525	< .0001
I$\sigma^2_{v_1}$	1	1.07	.301	< .0001	0.90	.342	< .0001	0.96	.3281	< .0001	0.26	.6079	< .0001
*J*K*	1	1.34	.247	< .0001	0.97	.324	< .0001	1.14	.2858	< .0001	0.00	.9670	< .0001
J$\sigma^2_{u_1}$	1	0.92	.338	< .0001	0.36	.549	< .0001	1.03	.3103	< .0001	0.70	.4016	< .0001
J$\sigma^2_{v_1}$	1	1.36	.243	< .0001	0.97	.324	< .0001	1.08	.2994	< .0001	< 0.00	.9672	< .0001
K$\sigma^2_{u_1}$	1	3.28	.070	< .0001	3.84	.050	< .0001	4.42	.0356	< .0001	3.28	.0701	< .0001
K$\sigma^2_{v_1}$	1	0.43	.512	< .0001	0.13	.722	< .0001	0.00	.9983	< .0001	0.87	.3517	< .0001
$\sigma^2_{u_1}$*$\sigma^2_{v_1}$	1	11.36	.001	.0001	0.13	.723	< .0001	0.00	.9752	< .0001	0.87	.3516	< .0001

Note. *$p < .001$. $\hat{\eta}^2$ reflects the magnitude of the main effects and/or interaction effects on the relative bias and mean squared error. $\hat{\eta}^2$ can take on values between 0 and 1 (small: $\hat{\eta}^2 \leq .02$, medium: $\hat{\eta}^2$ = .03-.25, and large: $\hat{\eta}^2 \geq .26$, Cohen, 1988). θ_{100} = immediate treatment effect, *I*, *J*, and *K* = number of measurements, cases, and studies respectively. Interactions between two independent variables are indicated with a * in between the two independent variables (e.g., θ_{100}*I indicates an interaction between the immediate treatment effect and the number of measurements).

Bias and *MSE* of the between-case variance of the treatment effect estimate. In addition to the estimated between-study variability in the treatment effect estimate, we were also interested in how accurate and precise the between-case variance of the treatment effect can be estimated using the three-level model. The relative bias [$F(1, 128004) = 0.35$, $p = .554$, $\hat{\eta}^2 < .0001$] and the *MSE* [$F(1, 128004) = 0.44$, $p = .508$, $\hat{\eta}^2 < .0001$] of the estimated between-case variance are independent of the analysis model. Also, no other design conditions have a significant influence on the relative bias and *MSE* as indicated in Table 4. The between-case variance estimate of the treatment effect does not exceeds the 5% criterion set by Boomsma and Hoogland (1998) except in the condition with a small amount of cases, a small amount of studies, a large amount of between study-variance and a small amount of between-case variance. Indeed the relative bias ranges from 0.0165% to 7.84%. This stands in contrast to the between-study variance estimates in which in almost all conditions biased estimates are obtained. In terms of the *MSE*, also smaller values are obtained compared to the between-study variance estimates. The range of the *MSE* is 0.0202 to 3.2793 (and is more than 13 times smaller than the *MSE* of the between-study variance). The conditions representing the largest values are all characterized by a small number of studies, a small number of cases and an unequal amount of between-case and between-study variance.

Simulation Study 2

In this second simulation study, we evaluated whether ignoring existing true covariance between the baseline level and the treatment effect has an effect on the overall treatment effect estimate (especially in terms of the standard error estimate and the coverage proportion of the 95% confidence interval). We were also interested whether Analysis Model 2 (in which covariance was modelled) resulted in more accurate (i.e., relative bias) and precise (i.e., *MSE*) variance component estimates.

Bias and *MSE* of the overall treatment effect estimate. Similar to the first simulation study, the relative bias [$F(1, 511538) = 0.03$, $p = .868$, $\hat{\eta}^2 < .0001$] and the *MSE* [$F(1, 511538) = 0.15$, $p = .709$, $\hat{\eta}^2 < .001$] appear to be independent of the analysis model. Compared to the first simulation study, we have additional design factors, namely the covariance (between the baseline level and the treatment effect) at the second and the third level but also these designs conditions have no significant effect on the bias (see Table 5). Similar to the first simulation study, the absolute bias ranges from -1.022×10^{-4} to 0.070 and the conditions representing the largest bias are characterized by 10 studies and 3 cases. A complete overview of the absolute bias per analysis model and design condition can be obtained by the first author.

Also in terms of the *MSE*, the same results were obtained as in Simulation Study 1: The *MSE* ranges from 0.066 to 1.166 and the largest *MSE* is found in the condition in which 10 studies, 3 cases, and 40 measurement occasions are included in combination with a large amount of between-case variance (i.e., $\sigma^2_{u_1} = 8$) and between-study variance (i.e., $\sigma^2_{u_1} = 8$). The same influencing design conditions (and interactions) are found as in Simulation Study 1 (see Table 5): The overall treatment effect can be estimated more precisely when the between-study and between-case variance are low. The *MSE* can further be reduced by including more units at level 2 and level 3 of the three-level model.

Table 5

Results Simulation Study 2 (i.e., Covariance is Generated) using Analysis Model 1 (no Covariance is Estimated). Evaluation of the Main Effects and Interaction Effects of Simulation Conditions on the Bias, Mean Squared Error, Relative Standard Error Bias, and Coverage Proportion of the 95% Confidence Interval of the Treatment Effect

		Bias			Mean Squared Error			Relative Standard Error Bias			Coverage Proportion 95% Confidence Interval		
Independent variable	*df*	*F*	*p*	$\hat{\eta}^2$	*F*	*p*	$\hat{\eta}^2$	*F*	*p*	$\hat{\eta}^2$	*F*	*p*	$\hat{\eta}^2$
θ_{100}	1	2.18	.139	<.0001	3.80	.051	<.0001	2.86	.0917	.0009	1.37	.2419	.0012
I	1	0.40	.528	<.0001	3.77	.052	<.0001	1.52	.2181	.0003	0.15	.7028	.0001
J	1	0.63	.426	<.0001	547.99*	<.0001	.0009	51.77*	<.0001	.0258	*38.60	<.0001	.0332
K	1	0.29	.591	<.0001	40874.00*	<.0001	.0675	169.22*	<.0001	.0853	0.17	.6832	.0001
$\sigma_{u_1}^2$	1	2.37	.690	.0001	623.21*	<.0001	.0031	46.11*	<.0001	.0686	*25.02	<.0001	.0645
$\sigma_{v_2}^2$	1	0.98	.430	<.0001	13283.80*	<.0001	.0658	287.23*	<.0001	.4356	*145.61	<.0001	.3753
θ_{100}*I	2	4.54	.330	<.0001	1.00	.318	<.0001	2.05	.1531	.0005	4.85	.0281	.0042
θ_{100}*J	1	0.30	.586	<.0001	1.18	.277	<.0001	0.07	.7918	-.0005	0.60	.4408	.0005
θ_{100}*K	1	2.23	.135	<.0001	1.99	.158	<.0001	0.00	.9698	-.0005	0.42	.5188	.0004
θ_{100}*$\sigma_{u_1}^2$	1	2.77	.400	<.0001	0.86	.460	<.0001	2.43	.0642	.0022	2.56	.0543	.0066
θ_{100}*$\sigma_{v_1}^2$	1	0.04	.989	<.0001	5.09	.002	<.0001	0.79	.4979	-.0003	1.20	.3083	.0031
*I*J	1	1.68	.195	<.0001	6.94	.008	<.0001	1.81	.1796	.0004	2.11	.1475	.0018
*I*K	1	0.49	.486	<.0001	6.10	.014	<.0001	1.32	.2512	.0002	1.51	.2200	.0013
I$\sigma_{u_1}^2$	1	0.39	.763	<.0001	0.97	.407	<.0001	2.64	.0491	.0025	2.51	.0582	.0065
I$\sigma_{v_1}^2$	1	0.43	.729	<.0001	4.31	.005	<.0001	2.06	.1053	.0016	1.74	.1582	.0045
*J*K	1	0.90	.342	<.0001	105.51*	<.0001	.0002	0.26	.6111	-.0004	0.29	.5879	.0003
J$\sigma_{u_1}^2$	1	2.33	.730	<.0001	96.78*	<.0001	.0005	1.31	.2697	.0005	0.53	.6606	.0014
J$\sigma_{v_1}^2$	1	0.07	.976	<.0001	16.60*	<.0001	.0001	13.44*	<.0001	.0189	5.95	.0005	.0153
K$\sigma_{u_1}^2$	1	1.36	.252	<.0001	3370.08*	<.0001	.0167	3.51	.0154	.0038	1.86	.1351	.0048
K$\sigma_{v_1}^2$	1	2.32	.730	<.0001	148.18*	<.0001	.0007	1.31	.2713	.0005	0.47	.7016	.0012
$\sigma_{u_1}^2$ *$\sigma_{v_1}^2$	1	1.35	.240	<.0001	25.09*	<.0001	.0004	21.61*	<.0001	.0941	*10.06	<.0001	.0778

Note. *p < .001. $\hat{\eta}^2$ reflects the magnitude of the main effects and/or interaction effects on the relative bias, mean squared error, relative standard error bias and coverage proportion of the 95% confidence interval. $\hat{\eta}^2$ can take on values between 0 and 1 (small: $\hat{\eta}^2 \leq .02$, medium: $\hat{\eta}^2$ = .03-.25, and large: $\hat{\eta}^2 \geq .26$, Cohen, 1988). θ_{100} = immediate treatment effect, *I*, *J*, and *K* = number of measurements, cases, and studies respectively. Interactions between two independent variables are indicated with a * in between the two independent variables (e.g., θ_{100}*I indicates an interaction between the immediate treatment effect and the number of measurements).

Standard error and coverage proportion of the 95% confidence interval of the overall treatment effect estimate. Similar to Simulation Study 1, the relative standard error bias [$F(1, 451) = 1.69$, $p = .194$, $\hat{\eta}^2 = .0009$] and the coverage proportion of the 95% confidence interval [$F(1, 451) = 1.91$, $p = .340$, $\hat{\eta}^2 = .0008$] are independent of the analysis model. However, different design conditions are found to have a statistically significant and medium effect on the relative standard error bias (see Table 5) compared to Simulation Study 1. There is a significant medium effect of the number of cases [$F(1, 451) = 51.77$, $p < .0001$, $\hat{\eta}^2 = .0263$], the number of studies [$F(1, 451) = 169.22$, $p < .0001$, $\hat{\eta}^2 = .0859$] and the covariance at the second level $F(3, 451) = 46.11$, $p < .0001$, $\hat{\eta}^2 = .0702$]. The covariance at the third level appears to have the largest impact [$F(3, 451) = 46.11$, $p < .0001$, $\hat{\eta}^2 = .4373$]. A statistically significant medium interaction effect is identified between the number of cases and the between-study covariance [$F(3, 451) = 13.44$, $p < .0001$, $\hat{\eta}^2 = .0205$] and between the covariance at the second level and the third level $F(9, 451) = 21.61$, $p < .0001$, $\hat{\eta}^2 = .0987$]. The large impact of the between-study covariance on the relative bias is illustrated in Figure 5. The relative standard error bias is smaller than 5% if the between-study covariance is large. If the between-study covariance is small, than the relative standard error bias increases if the between-study variance becomes larger (i.e., $\sigma^2_{u_1} = 8$). The results indicate that the relative bias standard error difference only exceeds the Hoogland and Boomsma's criterion of 10% in

conditions characterized by a large between-case covariance and a small between-study covariance (this represents only 4.89% of the conditions).

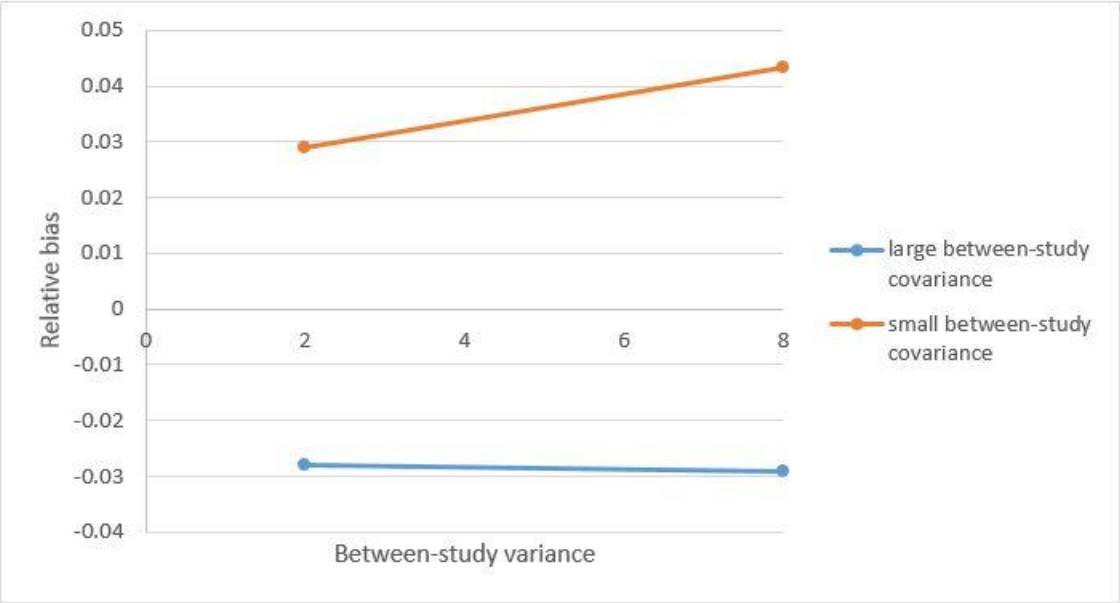

Figure 5. Graphical display of the influence of the between-study variance and the between-study covariance on the relative bias of the treatment effect in Simulation Study 2 (Covariance Estimated). The graphical display is for a subset for conditions as similar patterns where found for the other combination of conditions. Analysis Model 1 (i.e., no covariance is estimated) is used, the overall baseline level is set to zero, the number of measurements, cases and studies is set to 20, 3, and 10 respectively, the between-case variance is set to 0 and the between-case covariance is set to a small value.

In contrast to Simulation Study 1, we found statistically significant and medium to large effects of the design conditions on the coverage proportion of the 95% confidence interval (see Table 5). Indeed, the number of cases [$F(1, 451) = 38.60$, $p < .001$, $\hat{\eta}^2 = .0332$], the amount of between-case covariance [$F(3, 451) = 25.02$, $p < .001$, $\hat{\eta}^2 = .0645$] and the interaction between the amount of between-case covariance and the between-study covariance [$F(9, 451) = 10.06$, $p < .0001$, $\hat{\eta}^2 = .0778$] have a medium effect on the coverage proportion. In addition, the between-study covariance has a large effect [$F(1, 451) = 145.61$, $p < .0001$, $\hat{\eta}^2 = .3753$].

However, no problematic values for the coverage proportion of the 95% confidence interval were found as, similar to Simulation Study 1, all values lay between .93 and .98.

Bias and *MSE* of the between-study variance of the treatment effect estimate.

There is no significant influence of the used analysis model on the relative bias [$F(1, 511517) = 3.53$, $p = .060$, $\hat{\eta}^2 < .0001$] and the *MSE* [$F(1, 511517) = 1.93$, $p = .060$, $\hat{\eta}^2 < .0001$] of the between-study variance. Similar to Simulation Study 1, none of the design factors have a statistically significant large main effect on the relative bias. The same applies for the interaction effects. A complete overview of the main effects and the interaction effects accordingly with their effect sizes (i.e., $\hat{\eta}^2$) are displayed in Table 6. From this table we can deduce that the $\hat{\eta}^2$'s are smaller than .0001 for all main effects and interaction effects which indicates no major influences on the relative bias. The same applies for the *MSE* of the estimated between-study variance of the treatment effect as indicated in Table 6.

Table 6

Results Simulation Study 2 (i.e., Covariance is Generated) using Analysis Model 1 (no Covariance is Estimated). Evaluation of the Main Effects and Interaction Effects of Simulation Conditions on the Relative Bias and Mean Squared Error of the Between-Study Variance and Between-Case Variance Estimate

	Between-Study Variance							Between-Case Variance					
		Relative Bias			Mean Squared Error			Relative bias			Mean Squared Error		
Independent variable	*Df*	*F*	*p*	$\hat{\eta}^2$	*F*	*p*	$\hat{\eta}^2$	*F*	*p*	$\hat{\eta}^2$	*F*	*p*	$\hat{\eta}^2$
θ_{100}	1	0.64	0.424	<.0001	0.74	0.389	<.0001	1.70	0.1923	<.0001	1.58	0.2090	<.0001
I	1	0.71	0.400	<.0001	0.85	0.357	<.0001	0.56	0.4537	<.0001	0.64	0.4250	<.0001
J	1	0.35	0.552	<.0001	0.8	0.372	<.0001	0.36	0.5479	<.0001	0.60	0.4372	<.0001
K	1	2.29	0.130	<.0001	1.36	0.243	<.0001	1.49	0.2229	<.0001	1.34	0.2472	<.0001
$\sigma_{u_0u_1}$	3	2.64	0.048	<.0001	1.32	0.266	<.0001	2.71	0.0432	<.0001	1.80	0.1454	<.0001
$\sigma_{v_0v_1}$	3	1.78	0.149	<.0001	1.47	0.219	<.0001	1.62	0.1832	<.0001	1.01	0.3881	<.0001
θ_{100}*I	1	1.39	0.238	<.0001	1.49	0.223	<.0001	0.89	0.3464	<.0001	0.90	0.3441	<.0001
θ_{100}*J	1	1.03	0.311	<.0001	1.42	0.233	<.0001	0.63	0.4266	<.0001	0.86	0.3546	<.0001
θ_{100}*K	1	0.04	0.849	<.0001	0.37	0.541	<.0001	0.42	0.5161	<.0001	0.92	0.3370	<.0001
θ_{100}*$\sigma_{u_0u_1}$	3	1.87	0.133	<.0001	1.35	0.255	<.0001	2.64	0.0480	<.0001	1.89	0.1293	<.0001
θ_{100}*$\sigma_{v_0v_1}$	3	0.19	0.906	<.0001	0.44	0.722	<.0001	0.56	0.6440	<.0001	0.68	0.5650	<.0001
*I***J*	1	3.45	0.063	<.0001	1.93	0.165	<.0001	3.21	0.0733	<.0001	2.05	0.1518	<.0001
*I***K*	1	0.06	0.804	<.0001	0.45	0.503	<.0001	0.00	0.9785	<.0001	0.25	0.6165	<.0001
*I**$\sigma_{u_0u_1}$	3	2.16	0.090	<.0001	1.49	0.215	<.0001	1.01	0.3875	<.0001	0.84	0.4726	<.0001
*I**$\sigma_{v_0v_1}$	3	0.49	0.693	<.0001	0.65	0.584	<.0001	0.60	0.6179	<.0001	0.86	0.4613	<.0001
*J***K*	1	0	0.959	<.0001	0.41	0.520	<.0001	0.01	0.9342	<.0001	0.23	0.6307	<.0001
*J**$\sigma_{u_0u_1}$	3	1.34	0.259	<.0001	1.43	0.232	<.0001	0.72	0.5380	<.0001	0.80	0.4925	<.0001
*J**$\sigma_{v_0v_1}$	3	0.38	0.768	<.0001	0.61	0.611	<.0001	0.33	0.8019	<.0001	0.82	0.4812	<.0001
*K**$\sigma_{u_0u_1}$	3	1.05	0.370	<.0001	0.82	0.484	<.0001	1.52	0.2059	<.0001	1.07	0.3602	<.0001
*K**$\sigma_{v_0v_1}$	3	2.34	0.071	<.0001	1.66	0.173	<.0001	0.97	0.4077	<.0001	1.09	0.3517	<.0001
$\sigma_{u_0u_1}$*$\sigma_{v_0v_1}$	9	0.88	0.544	<.0001	0.91	0.516	<.0001	1.00	0.4412	<.0001	0.81	0.6054	<.0001

Note. All values for $\hat{\eta}^2$ are smaller than .0001 indicating a small effect ($\hat{\eta}^2 \leq .02$). *Main Effects and Interaction Effects of Simulation Conditions on the Relative Bias and Mean Squared Error of the Between-Study Variance Estimate and the Between-Case Variance Estimate*

We sorted the relative bias from small to large per analysis model. For Analysis Model 1, we found the relative bias ranging from 0.01% to 26.9%, whereas the relative bias ranges from 0.04×10^{-4} % to 23.5% for Analysis Model 2. The smallest values for the relative bias were found when the number of studies is 30 (independent of other design conditions). Indeed, values for the relative bias lower than the cut off criterion of 5% (i.e., Hoogland & Boomsma, 1998) all have 30 studies in common, which represents 43% out of the 256 design conditions for Analysis Model 1 and 43% out of the 256 conditions for Analysis Model 2. A similar amount of biased variance estimates are obtained for both models. From these results we conclude that there is no clear evidence that Analysis Model 2 results in more accurate between-study variance estimates (which confirms the ANOVA's conducted as primarily analysis).

Also in terms of the precision of the estimated between-study variance (i.e., *MSE*), there is no clear advantage of using Analysis Model 2 where covariance is modelled [$F(1, 511506) = 1.93$, $p = .165$, $\hat{\eta}^2 < .0001$]. Table 6 also illustrates that there are no statistically significant and large main effects or interaction effects on the *MSE* of the between-study variance of the overall treatment effect. The range for the *MSE* is similar across the two analysis models, namely between 0.145 and 13.306 for Analysis Model 1 and between 0.144 and 13.445 for Analysis Model 2. When sorting the *MSE* from small to large, we found that the conditions representing the smallest *MSE* all have 30 studies and a small amount of between-study variability in common. The largest *MSE* occurs when 10 studies with a large between-case variance is presented. A complete overview of the relative bias and the *MSE* per design condition and per analysis model can be requested by the first author.

Bias and *MSE* of the between-case variance of the treatment effect estimate.

Similar to the estimated between-study variance, the relative bias of the between-case variance is not significantly influenced by the used analysis model; $F(1, 511506) = 4.20$, $p =$

.040, $\hat{\eta}^2$ < .0001). No design conditions or interactions between conditions have a significant or large influence on the relative bias of the between-case variance of the treatment effect as indicated in Table 6. For Analysis Model 1, the relative bias of the between-case variance of the treatment effect ranges from -0.386% to 14.7% and for Analysis model 2 from -5.750×10^{-4}% to -7.7. This means that the range of the relative bias using Analysis Model 1 is almost twice as large compared to Analysis Model 2. We found that in 92.97% of the conditions the relative bias is lower than the 5% cut off criterion if Analysis Model 2 is used whereas this is only 61.33% if Analysis Model 1 is used. Although no statistically significant and large effect of the analysis model on the relative bias is found (see Table 6), the relative bias is smaller in all conditions using Analysis Model 2. The largest difference in relative bias occurs in conditions having a small number of level 1, level 2 and level 3 units with a small amount of between-case and between-study variance. The largest difference equals 10%.

In terms of the *MSE* of the estimated between-case variance of the treatment effect, no statistically significant and large effect of the analysis model; $F(1, 511506) = 2.05$, $p = .209$ $\hat{\eta}^2 < .0001$) was found. The same accounts for the design conditions and interactions (see Table 6). The largest differences in *MSE* are observed in the conditions representing a large amount of between-study and between-case variance, accordingly with large values for the covariance. The *MSE* using Analysis Model 2 is smaller than or equal to the *MSE* using Analysis Model 1 in 76% of the conditions.

Bias and *MSE* of between-case covariance and between-study covariance estimates. In the second simulation study where covariance is modelled and estimated (i.e., Analysis Model 2), the preliminary ANOVA indicates that none of the design conditions have a statistically significant or large influence on the relative bias and MSE of the estimated between-study covariance as indicated in Table 7. The estimated between-study covariance is extremely biased in all conditions and ranges from 169% to 206% (see Table 10 for a

complete overview of the relative bias per design condition). None of the design conditions succeeds in reducing the bias. Also the *MSE* appears to be rather high having values ranging from 1.13 to 124.22. This indicates that the three-level model is unable to estimate the between-study covariance accurately and precisely when there is in reality non-zero between-study covariance. Especially in the condition representing a large amount of between-study covariance extremely biased and imprecise estimates are obtained.

Table 7

Results Simulation Study 2 (i.e., Covariance is Generated) using Analysis Model 1 (no Covariance is Estimated). Evaluation of the Main Effects and Interaction Effects of Simulation Conditions on the Relative Bias and Mean Squared Error of the Between-Study Covariance and the Between-Case Covariance Estimate

	Between-Study Covariance							Between-Case Covariance					
		Relative Bias			Mean Squared Error			Relative Bias			Mean Squared Error		
Independent Variable	*df*	*F*	*p*	$\hat{\eta}^2$	*F*	*p*	$\hat{\eta}^2$	*F*	*p*	$\hat{\eta}^2$	*F*	*p*	$\hat{\eta}^2$
θ_{100}	1	1.52	0.2181	<.0001	2.03	0.154	<.0001	1.40	0.236	<.0001	0.90	0.3414	<.0001
I	1	1.17	0.2791	<.0001	1.86	0.173	<.0001	0.69	0.405	<.0001	0.51	0.4759	<.0001
J	1	1.45	0.2289	<.0001	1.85	0.173	<.0001	0.72	0.397	<.0001	0.51	0.4762	<.0001
K	1	0.92	0.3374	<.0001	0.06	0.808	<.0001	1.25	0.264	<.0001	1.45	0.2292	<.0001
$\sigma_{u_0u_1}$	3	1.87	0.1320	<.0001	2.08	0.101	<.0001	2.70	0.044	<.0001	1.59	0.1900	<.0001
$\sigma_{v_0v_1}$	3	0.83	0.4744	<.0001	0.73	0.532	<.0001	1.38	0.245	<.0001	1.40	0.2413	<.0001
θ_{100}*I	1	1.22	0.2684	<.0001	1.86	0.172	<.0001	0.77	0.380	<.0001	1.50	0.2213	<.0001
θ_{100}*J	1	1.28	0.2583	<.0001	1.87	0.172	<.0001	0.81	0.369	<.0001	1.50	0.2213	<.0001
θ_{100}*K	1	0.31	0.5766	<.0001	0.05	0.831	<.0001	0.15	0.699	<.0001	0.28	0.5960	<.0001
θ_{100}*$\sigma_{u_0u_1}$	3	1.78	0.1480	<.0001	2.06	0.104	<.0001	2.09	0.099	<.0001	1.89	0.1293	<.0001
θ_{100}*$\sigma_{v_0v_1}$	3	0.49	0.6874	<.0001	0.69	0.559	<.0001	0.51	0.678	<.0001	0.33	0.8026	<.0001
*I*J*	1	2.63	0.1046	<.0001	2.11	0.146	<.0001	3.69	0.055	<.0001	2.64	0.1044	<.0001
*I*K*	1	0.13	0.7196	<.0001	0.02	0.880	<.0001	0.00	0.960	<.0001	0.08	0.7707	<.0001
I$\sigma_{u_0u_1}$	3	1.66	0.1737	<.0001	1.88	0.130	<.0001	1.27	0.282	<.0001	1.30	0.2719	<.0001
I$\sigma_{v_0v_1}$	3	0.77	0.5103	<.0001	0.65	0.581	<.0001	0.59	0.623	<.0001	0.42	0.7391	<.0001
*J*K*	1	0.21	0.6461	<.0001	0.02	0.880	<.0001	0.00	0.951	<.0001	0.09	0.7704	<.0001
J$\sigma_{u_0u_1}$	3	1.77	0.1510	<.0001	1.88	0.130	<.0001	1.25	0.289	<.0001	1.30	0.2717	<.0001
J$\sigma_{v_0v_1}$	3	0.85	0.4638	<.0001	0.65	0.581	<.0001	0.58	0.627	<.0001	0.42	0.7391	<.0001
K$\sigma_{u_0u_1}$	3	1.41	0.2388	<.0001	1.28	0.278	<.0001	1.44	0.229	<.0001	1.61	0.1843	<.0001
K$\sigma_{v_0v_1}$	3	0.49	0.6899	<.0001	0.05	0.984	<.0001	0.76	0.516	<.0001	0.73	0.5337	<.0001
$\sigma_{u_0u_1}$*$\sigma_{v_0v_1}$	9	0.80	0.6165	<.0001	0.71	0.701	<.0001	0.97	0.460	<.0001	0.74	0.6716	<.0001

Note. $\hat{\eta}^2$ reflects the magnitude of the main effects and/or interaction effects on the relative bias and mean squared error. $\hat{\eta}^2$ can take on values between 0 and 1 (small: $\hat{\eta}^2 \leq .02$, medium: $\hat{\eta}^2$ = .03-.25, and large: $\hat{\eta}^2 \geq .26$, Cohen, 1988). θ_{100} = immediate treatment effect, *I*, *J*, and *K* = number of measurements, cases, and studies respectively. $\sigma_{u_0u_1}$, $\sigma_{v_0v_1}$ indicate the between-case covariance and between-study covariance respectively. $\sigma^2_{u_1}$ and $\sigma^2_{v_1}$ indicate the between-case

For the estimated between-case covariance no statistically significant or large effects of the design conditions on the relative bias and the *MSE* were found as indicated in Table 7. The relative bias of the between-case covariance ranges from 0.01% to 14.74%. The between-case covariance estimates are unbiased (i.e., smaller than 5%) if there are at least 30 studies included (independent of other conditions). In case only 10 studies are available for synthesis, then the between-case covariance estimate is unbiased if each study includes at least 7 cases. In these conditions, the between-case covariance is most precisely estimated (indicated by smaller values for the *MSE*). The *MSE* ranges from 0.011 to 2.18 and the largest values coincidence with the conditions having the largest relative bias (i.e., 10 studies and 3 cases). This indicates that the three-level model is appropriate to estimate the between-case covariance accurately and precisely when there is in reality non-zero between-case covariance at least when there are 7 cases within each study included. When only 3 cases are included, than combining at least 30 studies is necessary.

In sum, an interesting finding is that the between-study covariance is not well estimated in terms of accuracy and precision using the three-level model and therefore can induce biased between-study variance estimates. In contrast, the between-case covariance estimates tend to be unbiased and well estimated. This explains why the between-case variances are unbiased in most conditions and more precisely estimated compared to the between-study variance estimates. These results indicate why Analysis Model 2 does not outperform Analysis Model 1 when covariance is generated.

Empirical Illustration

We use the meta-analysis of single-case studies conducted by Denis et al. (2011). They collected studies where the effectiveness of a treatment for self-injurious behavior in people with profound intellectual disabilities was investigated. In particular, 18 studies were collected where non-aversive, non-intrusive forms of reinforcement were examined. The single-case

experiments were coded as AB phase designs. We analyzed the data by modeling possible covariance between the regression coefficients (i.e., Analysis Model 2) and ignoring possible covariance between the regression coefficients (i.e., Analysis Model 1).

The results indicate that, as expected, the ignorance of existent covariance has no large effect on the estimated treatment effects. The overall treatment effect estimates are statistically significant at the .01 level and equal -2.43 [$t(18.2) = –.67$, $p < .01$] for Analysis Model 1 and -2.41 [$t(18.2) = -4.71$, $p < .01$] for Analysis Model 2. From these analyses we can conclude that the treatment is effective to reduce self-injurious behavior.

The estimated between-study variance of the treatment effect is slightly larger when Analysis Model 1 is used ($\hat{\sigma}^2_{v_1} = 4.34$, $Z = 2.59$, $p = < .01$) in comparison to Analysis Model 2 ($\hat{\sigma}^2_{v_1} = 3.96$, $Z = 2.59$, $p = < .01$). We also identified a difference between the estimated between-case variance of the treatment effect, which equals 0.54 [$Z = 1.49$, $p = .07$], for Analysis Model 1 and 1.02 [$Z = 1.71$, $p = .04$] for Analysis Model 2. The estimated covariance, in Analysis Model 2 between the baseline level and the treatment effect equals, -0.92 [$Z = -1.80$, $p = .07$] and -3.09 [$Z = –1.62$, $p = .10$] at level 2 and level 3 respectively. This means that a large estimated baseline level at level 2 and level 3 go together with a small estimated treatment effect.

This empirical example indicates that the fixed effect estimates are independent of the analysis model (i.e., the treatment is found to be effective to reduce self-injurious behavior). For the variance components estimates, we can feel confident in the estimate of the between-case variance and covariance. However, caution is needed when interpreting the between-study variance and covariance as biased and less precise estimates can be obtained, especially when less than 30 studies are included (as is the case in this empirical example). In sum, if the research interest lies in the between-case variance estimates and the treatment effect estimate, the multilevel model (either Analysis Model 1 or Analysis Model 2) can be applied to the dataset of Denis et al. (2011). If the research interest also lies in the between-case covariance estimate

only Analysis Model 2 can be used. In this empirical example, the multilevel model is not appropriate to obtain accurate and reliable between study (co)variances as less than 30 studies are included. If the research interest lies in the between-study variance we advised to search in the literature for similar focused single-case studies to include in the synthesis.

Discussion

General Conclusion

The main purpose of this study was to evaluate the consequences of misspecifying the between-case and between-study covariance matrix on the estimation of the treatment effects, and their corresponding mean squared error, standard errors, coverage proportion of the 95% confidence interval and variance and covariance components. Because it is not always obvious how to define the covariance matrix, it is important to examine the degree to which the treatment effect estimates and the variance estimates are sensitive to changes in the specification of the covariance matrix.

In contexts of SCD data, only the effects of level-1 residuals' covariance misspecification have been studied before for a two-level model (Ferron, et al. , 2009). Previous methodological work devoted to misspecification of the (co)variance matrix has been conducted in contexts other than multilevel modeling of SCDs (Berkhof & Kampen, 2004; Kwok et al., 2007; Jahng, 2008; Moerbeek, 2004; Singer & Willett; Van den Noortgate & Onghena, 2005). Level-2 and level-3 covariance misspecification issues in the SCD three-level modeling framework have not yet been investigated. Indeed, previous research concerning the three-level model for SCD data focused on the estimate of the overall treatment effect, and of the between-case and between-study variance of the overall treatment effect. The between-case and between-study residuals were each assumed to be independently, identically, and normally distributed with mean zero and homogeneous variance, and thus a diagonal covariance structure was assumed. To date, no research has focused on the consequences of ignoring truly non-zero covariances in the context of multilevel modeling of SCD data. However, this is urgent as it

provides a more complete understanding of misspecification issues in contexts of three-level modeling of SCD data.

In Simulation Study 1, we compared the scenario in which covariance is not simulated and not estimated (i.e., Analysis Model 1) with the scenario in which covariance is not simulated, but is estimated (i.e., Analysis Model 2). No statistically significant and large differences between Analysis Model 1 and Analysis Model 2 were found in terms of both the overall treatment effect estimate and variance components estimates. The overall treatment effect estimate as well as the between-case variance is estimated accurately and precisely. However, the between-study variance estimates are biased and not precisely estimated when the between-case variance and between-study variance have different values and when less than 30 studies are included. This confirms previous research devoted to multilevel modeling of single cases (Moeyaert et al. 2013a, 2013b; Owens & Ferron, 2012). In sum, if there is in reality no covariance, using Analysis Model 1 or Analysis Model 2 does not make a difference both in terms of the overall treatment effect or variance estimates. Caution is needed when estimating the between-study variance because in certain conditions biased and less precise estimates are obtained.

In the second simulation study, we compared the condition where covariance is simulated, but ignored in the analysis (i.e., Analysis Model 1) with the scenario where covariance is simulated and estimated in the analysis (i.e., Analysis Model 2). The overall treatment effect is equally well estimated using both analysis models. Surprisingly, Analysis Model 2 did not outperform Analysis Model 1 for the variance component estimates. Also in this simulation study, biased and less precise between-study (co)variance estimates are obtained in contrast to the between-case (co)variance estimates both when Analysis Model 1 and Analysis Model 2 are applied. We expected biased variance estimates for Analysis Model 1, because the model is misspecified by ignoring true existing covariance. However, also the correctly specified model (i.e., Analysis Model 2) resulted in biased and not well estimated

between-study variance estimates. This can be cause by the extremely biased between-study covariance estimates.

As a consequence, we advise researchers interested in estimating the overall treatment effect across cases and across studies (in similar contexts as the conditions included in this simulation study) to use either the analysis model including or not including covariance. However, when the research interest lies in the (co)variance components estimates, either Analysis Model 1 or Analysis Model 2 result in unbiased between-case (co) variance estimates. Caution is warranted when the between-study (co) variances are of interest.

Further research is needed in this context, because we expected that the model in which covariance was modelled would outperform the model in which covariance was ignored, but this was not the case. Also the between-study covariance estimates were biased and not well estimated. Therefore, there is a need to search for alternative estimation procedures such as bootstrapping and Bayesian estimation recommended in contexts dealing with small sample sizes as is the case in single-case research.

Limitations

As with any simulation study, one of the major potential limitations of this study is the generalizability of the findings. Further research is needed for the applicability of current findings to a broader range of conditions. We partly addressed this limitation by including realistic conditions based on several re-analyses of meta-analyses. The conditions are quite representative for the research field of single-case experiments in educational settings.

In current research, we only investigated the basic MBD including one predictor at the first level (i.e., dummy coded variable indicating the treatment effect). We excluded models with multiple predictors at level 2 and level 3, models using unbalanced data, non-linear models, reversal and alternating designs, and other complex models.

In this study, we only included covariance at the second and third level, which means that we ignored possible autocorrelation at the first level. The issue of autocorrelation itself

deserves separate research and is beyond the scope of this paper (Baek & Ferron, 2013). In Simulation Study 2, we generated covariance simultaneously at level 2 and level 3 (taking on different values and crossing these values) and analyzed the generated datasets including or not including covariance at level 2 and level 3. We did not include an analysis model in which covariance was only modelled at one level and ignored at the other level, which would be interesting for further research.

The combination of SCD data over studies may be difficult if studies are too different. Studies may for instance differ in measuring the treatment effect. We can handle this by the inclusion of covariates indicating certain study and even case characteristics to model this heterogeneity. Another possibility is standardizing the data or using a multivariate three-level model.

Other approaches to estimate the treatment effects and variances in these treatment effects when the variance structures are misspecified should be considered in future research, such as the sandwich estimator (i.e., cluster-robust or Huber estimators). Even when the covariance matrices are misspecified, the sandwich estimator is asymptotically consistent (Raudenbush & Bryck, 2002; Hedges, Tipton, & Johnson, 2010). It would be a useful contribution to compare the standard errors and coverage proportion of the 95% confidence intervals constructed with the sandwich estimator to those constructed using the model-based estimators in the misspecified model.

Furthermore, the misspecification of the covariance matrix is only one aspect to test the robustness of the three-level modeling approach. Further research is needed to evaluate other issues such as non-normal data and not identical distributed errors. However, as no previous research yet focused on misspecification of the covariance matrix in contexts of multilevel modeling of single-cases, this study provides some important insights. We advise single-case researchers to consider use of the three-level model, either modeling or ignoring covariance, when the research interest lies in the fixed effects or the between-case variance. If the research

interest lies in the between-case covariance the three-level model taking into account covariance can be used. The three-level models appear to be less appropriate to estimate between-study (co)variance, especially when there are less than 30 studies included.

References

Alen, E., Grietens. H., & Van den Noortgate. W. (2009). *Meta-analysis of single-case studies: An illustration for the treatment of anxiety disorders.* Unpublished manuscript.

Baek, E., & Ferron, J. M. (2013). Multilevel models for multiple-baseline data: Modeling across participant variation in autocorrelation and residual variance. *Behavior Research Methods, 45,* 65-74.

Barlow, D. H., Nock, M. K., & Hersen, M. (2009). *Single case experimental designs: Strategies for studying behavior change* (3rd ed.). Boston, MA: Allyn & Bacon.

Berkhof, J., Kampen, J. K. (2004). Asymptotic effect of misspecification in the random part of the multilevel model. *Journal of Educational and Behavioral Statistics, 29,* 201–218.

Hoogland, J.J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research, 26,* 329–367.

Denis, J., Van den Noortgate, W., & Maes, B. (2011). Self-injurious behavior in people with profound intellectual disabilities: A meta-analysis of single-case studies. *Research in Developmental Disabilities, 32,* 911-923.

Farmer, J., Owens, C. M., Ferron, J.M., & Allsopp, D. (2010*). A review of social science single-case meta-analyses.* Manuscript in preparation.

Ferron, J., & Scott, H. (2005). Multiple baseline designs. In B. Everitt & D. Howell (Eds.). *Encyclopedia of Behavioral Statistics* (Vol. 3. pp. 1306-1309). West Sussex. UK: Wiley & Sons Ltd.

Ferron, J. M., Bell, B. A., Hess, M. F., Rendina-Gobioff, G., & Hibbard. S. T. (2009). Making treatment effect inferences from multiple-baseline data: the utility of multilevel modeling approaches. *Behavior Research Methods, 41,* 372-384.

Ferron, J. M., Farmer, J. L., & Owens, C. M. (2010). Estimating individual treatment effects from multiple-baseline data: A Monte Carlo study of multilvel-modeling approaches. *Behavior Research Methods, 42*, 930-943.

Hedges, L. V, Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*(1), 39–65. doi:10.1002/jrsm.5

Heyvaert, M., Maes, B., Van den Noortgate, W., Kuppens, S., & Onghena, P. (2012). A multilevel meta-analysis of single-case and small-n research on interventions for reducing challenging behavior in persons with intellectual disabilities. *Research in Developmental Disabilities, 33* (2), 766-780.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling. An overview and a meta-analysis. *Sociological Methods & Research*, *26,* 329-367.

Jahng, S. (2008). *Multilevel models for intensive longitudinal data with heterogeneous error structure covariance transformation and variance function models.* (Doctoral dissertation). University of Missouri, Columbia.

Kazdin, A.E. (2011). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). New York: Oxford University Press.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics, 53*, 983-997.

Kinugasa. T., Cerin. E., & Hooper, S. (2004). Single-Subject Research Designs and Data Analyses for Assessing Elite Athletes' Conditioning. *Sports Medecine, 34,* 1035-1050.

Koehler, M.J., Levin, J.R. (2000). RegRand: Statistical software for the multiple-baseline design. *Behavior Research Methods, Instruments, and Computers, 32(*2), 367-71.

Kokina, A., & Kern, L. (2010). Social story interventions for students with autism spectrum disorders: a meta-analysis. *Journal of Autism and Developmental Disorders, 40*, 812-826.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M. , & Shadish, W. R. (2010). *Single-case designs technical documentation*. Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

Kromrey, J.D., & Foster-Johnson, L. (1996). Determining the efficacy of intervention: The use of effect sizes for data analysis in single-subject research. *The journal of Experimental Education, 65*, 79-96.

Kwok, O., West, S.G., Green, S.B. (2007). The impact of misspecifying the within-subject covariance structure in multiwave longitudinal multilevel models: a monte carlo simulation study. *Multivariate Behavioral Research, 42,* 557-592.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS© system for mixed models* (2nd ed.). Cary, NC: SAS Institute Inc.

Moerbeek M. *(2004).* The consequence of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research, 39,* 129-149.

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, T., & Van den Noortgate, W. (2013a). Three-level analysis of single-case experimental data: Empirical validation. *The Journal of Experimental Education*, *82*, 1-21.

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S. N., & Van den Noortgate, W. (2013b). *Three-level analysis of standardized single-case experimental data: Empirical validation.* Multivariate Behavior Research, *48*, 719-748

National Research Council (1992). *Combining information. Statistical issues and opportunities for research.* Washington, D.C.: National Academy Press.

Odom, S. L., Brantlinger, E., Gersten, R., Horner, R. H., Thompson, B., Harris, K. R. (2005). Research in special education: Scientific methods and evidence-based practices. *Exceptional Children, 71*(2), 137—148.

Onghena, P. (2005). Single-case designs. In B. Everitt & D. Howell (Eds.). *Encyclopedia of statistics in behavioral science* (Vol. 4. pp. 1850-1854). Chichester: Wiley.

Onghena, P., & Edgington, E. S. (2005). Customization of pain treatments: Single-case design and analysis. *Clinical Journal of Pain. 21*, 56-68.

Owens, C. M., & Ferron, J. M. (2012). Synthesizing single-case studies: A Monte Carlo examination of a three-level meta-analytic model. *Behavior Research Methods, 44*, 795-805.

Petit-Bois, M., Beak, E. K., Ferron, J. M. Consequences of misspecification of growth trajectories when meta-analyzing single-case data using a three-level model. Paper presented at the American Educational Research Association conference, Vancouver, British Columbia, Canada, 13-17 April 2012.

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical Linear Models: Applications and Data Analysis Methods (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods, 43*, 971-980.

Shogren, K. A., Fagella-Luby, M. N., Bae, J. S., & Wehmeyer, M. L. (2004). The effect of choice-making as an intervention for problem behavior. *Journal of Positive Behavior Interventions, 6*(4), 228-237.

Singer, J.D., & Willett, J.B. (2003). *Applied longitudinal data analysis: Model change and event occurrence.* Oxford: Oxford University Press.

Van den Noortgate, W., Opdenakker, M. & Onghena, P. (2005). The effects of ignoring a level in mutlilevel analysis. *School effectiveness and School Improvement, 16,* 281-303.

Wang, S., Cui, Y., & Parrila, R. (2011). Examining the effectiveness of peer-mediated and video-modeling social skills interventions for children with autism spectrum disorders: a meta-analysis in single-case research using HLM. *Research in Autism Spectrum Disorders, 5*, 562-569.